MODELO ANALÍTICO PARA LA PREDICCIÓN DE LA DESERCIÓN ESTUDIANTIL A NIVEL DE PREGRADO EN LA UNIVERSIDAD AUTÓNOMA DEL CARIBE

LUIS JOSÉ POLO AHUMADA
DARLING JOHANA MEDINA REYES

UNIVERSIDAD AUTÓNOMA DEL CARIBE
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA MECATRÓNICA
BARRANQUILLA - COLOMBIA
2022

MODELO ANALÍTICO PARA LA PREDICCIÓN DE LA DESERCIÓN ESTUDIANTIL A NIVEL DE PREGRADO EN LA UNIVERSIDAD AUTÓNOMA DEL CARIBE

LUIS JOSÉ POLO AHUMADA
DARLING JOHANA MEDINA REYES

Trabajo de grado presentado para optar al título de

Ingeniero Mecatrónico

ASESORES DISCIPLINARES:

Ing. Jair Villanueva, PhD
Ing. Carlos Díaz, Ms

UNIVERSIDAD AUTÓNOMA DEL CARIBE

FACULTAD DE INGENIERÍA

PROGRAMA DE INGENIERÍA MECATRÓNICA

BARRANQUILLA - COLOMBIA

2022

Nota de aceptación

_____

_____

_____

_____

_____

_____

_____

Firma del jurado 1

_____

Firma del jurado 2

DEDICATORIA

Queremos agradecer a Dios por permitirnos lograr una meta tan importante en nuestras vidas.

Agradecer a la Universidad Autónoma del Caribe, su cuerpo docente y de apoyo por brindarnos la oportunidad de hacer parte de este maravilloso centro educativo y de apoyarnos en la ejecución de este proyecto.

A nuestros asesores de proyecto de grado Jair Villanueva Padilla y Carlos Díaz Sáenz tutores, a nuestros evaluadores Saúl Pérez, Jean Coll y Kelvin Beleño y al programa de Permanencia Académica en cabeza de Sol Barraza, los cuales nos orientaron en la consecución de este proyecto.

A nuestros compañeros de pregrado por su apoyo, confianza, por compartir sus conocimientos y fortalecer los lazos de amistad.

A nuestras familias por ser un pilar fundamental en todo este proceso y apoyarnos en cada instante.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

GLOSARIO

Machine Learing: es una disciplina del campo de la Inteligencia Artificial que, a través de algoritmos, dota a los ordenadores de la capacidad de identificar patrones en datos masivos y elaborar predicciones (análisis predictivo).

Business Intelligence: (BI por sus siglas en inglés) hace referencia al uso de estrategias y herramientas que sirven para transformar información en conocimiento, con el objetivo de mejorar el proceso de toma de decisiones en una empresa.

Aprendizaje automático: El aprendizaje automático (ML) es una subcategoría de inteligencia artificial que se refiere al proceso por el cual los PC desarrollan el reconocimiento de patrones o la capacidad de aprender continuamente y realizar predicciones basadas en datos, tras lo cual realizan ajustes sin haber sido programados específicamente para ello.

Deserción estudiantil: Puede entenderse como el abandono del sistema escolar por parte de los estudiantes, provocado por la combinación de factores que se generan tanto al interior del sistema como en contextos de tipo social, familiar, individual y del entorno.

Analítica de datos: permite a las organizaciones analizar todos sus datos (en tiempo real, históricos, no estructurados, estructurados, cualitativos) para identificar patrones y generar conocimientos para informar y, en algunos casos, automatizar decisiones, conectando la inteligencia y la acción.

Predicción: que **anticipa aquello que, supuestamente, va a suceder**. Se puede predecir algo a partir de conocimientos científicos, relevaciones de algún tipo, hipótesis o indicios.

Educación superior: está conformada por los programas educativos “posteriores a la enseñanza secundaria, impartidos por universidades u otros establecimientos que estén habilitados como instituciones de enseñanza superior por las autoridades competentes del país y/o sistemas reconocidos de homologación.

Árbol de decisión: es un mapa de los posibles resultados de una serie de decisiones relacionadas. Permite que un individuo o una organización comparen posibles acciones entre sí según sus costos, probabilidades y beneficios.

Base de datos: se encarga no solo de almacenar datos, sino también de conectarlos entre sí en una unidad lógica.

Regresión logística: es una técnica de análisis de datos que utiliza las matemáticas para encontrar las relaciones entre dos factores de datos. Luego, utiliza esta relación para predecir el valor de uno de esos factores basándose en el otro.

RESUMEN

El objetivo principal de este proyecto de grado obedece al cumplimiento de una investigación la cual consistió en la creación de un modelo analítico para la predicción de la deserción estudiantil a nivel de pregrado en la Universidad Autónoma del Caribe, A partir de la implementación de este modelo, la Universidad Autónoma del Caribe está en la capacidad de implementar estrategias que permitan disminuir la tasa de deserción de estudiantes de pregrado, esta investigación se abordó a partir del análisis de diferentes factores socioeconómicos y académicos, además requirió de la ejecución de una serie de fases: caracterización, experimentación, desarrollo y evaluación. Durante las diferentes fases se construyeron conjunto de datos (dataset), se utilizaron técnicas de aprendizajes (machine learning), así como la utilización de herramientas como Power BI, cuyos resultados permiten segmentar la población estudiantil en riesgo de deserción a través de tres niveles: riesgo bajo, riesgo medio y riesgo alto, lo cual permitirá focalizar estrategias de permanencia académica sobre estudiantes que se encuentren categorizados principalmente en riesgo alto de deserción.

Palabras clave: Machine Learning, Inteligencia de negocio.

ABSTRACT

The present degree work is due to the fulfillment of an investigation which consists of the creation of an analytical model for the prediction of the dropout of undergraduate students at the Autonomous University of the Caribbean, From the implementation of this model, Universidad Autónoma del Caribe is capable of implementing strategies to reduce the dropout rate of undergraduate students. This research was approached from the analysis of different socioeconomic and academic factors, and also required the execution of a series of phases: characterization, experimentation, development and evaluation. During the different phases, datasets were built, learning techniques (machine learning) were used, as well as the use of tools such as (power model, Power BI), whose results were analyzed for a subsequent conclusion and recommendations to be used. aimed at a solution and improvement.

Keywords: Machine Learning, Business Intelligence.

INTRODUCCIÓN

La siguiente investigación reúne los conceptos sobre la deserción estudiantil el cual, es un fenómeno que afecta a muchos actores de la sociedad, en especial a la población juvenil sobre la cual se crean grandes desilusiones, al ingresar al sistema de educación superior y por diversos motivos no logran el tan anhelado objetivo.

Esto deja en evidencia las fallas que existen en el sistema de educación superior actual, ya que son muy pocas las estrategias que ayuden a mitigar los altos índices de deserción en pro de mantener el mayor número de estudiantes posibles que ingresan a la educación superior, esto sin detrimento del proceso formativo.

Lo anterior mencionado da como resultado un modelo analítico para la predicción de la deserción estudiantil en la Universidad Autónoma del Caribe, aportando estrategias implementables, que permitan disminuir la tasa de deserción estudiantil a nivel de pregrado, lo que representaría un aumento en los recaudos por concepto de matrículas.

Este proyecto fue realizado con fines académicos, para contribuir con aportes tecnológicos, asimismo de conocimientos tanto para los investigadores, como para la Universidad Autónoma del Caribe. Para el desarrollo de este se utilizó un método de investigación analítico predictivo, usando técnicas de Machine Learning y herramientas como power BI, para la recopilación de datos, procesamiento de datos, desarrollo de modelo predictivo y la posterior visualización de la información.

La finalidad del proyecto estipula el cumplimiento de los objetivos específicos los cuales se estipulo la identificación y documentación de las variables socioeconómicas y académicas de los estudiantes, y las técnicas Machine Learning como apoyo a la construcción del conjunto de datos, así como la

implementación de un modelo predictivo para predecir la probabilidad estudiantil a nivel de pregrado en la UAC, de desertar en su programa académico, por ultimo evaluar el modelo predictivo propuesto a partir de la implementación, validados mediante el análisis de la metica de calidad.

Asimismo, el proyecto está distribuido en seis capítulos, en el primer capítulo se realizó el planteamiento del problema, el cual consta de antecedentes, formulación del problema y la justificación del alcance, lo que arroja como consecuencia la formulación de la pregunta problema ¿Qué herramienta tecnológica, basada en técnicas de Machine Learning, permitiría mejorar los niveles de deserción estudiantil a nivel de pregrado en la Universidad Autónoma del Caribe?

En el segundo capítulo, se presenta el objetivo general de la investigación "Desarrollar un modelo analítico que permita predecir la deserción estudiantil a nivel de pregrado en la Universidad Autónoma del Caribe" y los objetivos específicos para alcanzar este.
En el tercer capítulo se muestra el marco de referencia donde se evidencia el estado del arte, el cual se podrá observar los proyectos realizados anteriormente, los cuales fueron base del proyecto desarrollado, también cuenta con el marco teórico con los fundamentos necesarios para la investigación.

En el cuarto capítulo se muestra el procedimiento metodológico el cual se basó el grupo investigador para acabo el proyecto y posterior cumplimiento de los objetivos, asimismo se observa el cronograma y plan de trabajo. En el quinto capítulo se expone el presupuesto general, personal científico y de apoyo. Por último, en el sexto se evidencia el modelo analítico, la recolección de datos, análisis de resultado y manual de usuario.

1. PLANTEAMIENTO DEL PROBLEMA

Basados en los lineamientos definidos por el MEN y a partir del análisis de la información extraída del SPADIES, la Universidad Autónoma del Caribe, ha diseñado e implementado diferentes estrategias, asimismo ha creado varios programas, estos direccionados a promover la permanencia estudiantil, tales como: Modelo de Gestión de Permanencia y Graduación, Modelo de Seguimiento Académico Integral por Fases.

Todas estas estrategias definidas en la Política Institucional de Permanencia Académica con Calidad y Excelencia PACE. Los anteriores programas y estrategias mencionadas han incidido positivamente en la disminución de índices de deserción estudiantil a nivel de pregrado en la UAC, gracias a la identificación de factores de riesgo.

De acuerdo con el conversatorio “Perspectivas de Permanencia Académica” realizado por los docentes de la UAC, han surgido nuevos factores de riesgo, que afectan directamente el comportamiento del fenómeno de deserción estudiantil y, si bien las estrategias implementadas han logrado disminuir estos índices, estas no han sido totalmente eficaces, para disminuir de manera significativa los porcentajes de deserción y, además, que se encuentren alineados con los requerimientos del MEN.

A partir de los anterior, nace la necesidad de desarrollar un modelo analítico para la predicción de la deserción estudiantil a nivel de pregrado en la Universidad Autónoma del Caribe, con el cual se buscó una solución que permitirá a la UAC aplicar de manera más efectiva las estrategias para disminuir las tasas de deserción estudiantil, Lo anterior, permite centralizar las variables y los factores objeto de análisis que favorecen la toma de decisiones en pro de mitigar el fenómeno de la deserción estudiantil a nivel de pregrado.

1.1. ANTECEDENTES

El fenómeno de la deserción estudiantil conlleva consigo una dificultad al momento de implementar los procesos que se enfocan en ampliar la cobertura del sistema educativo de educación superior, lo que da como resultado postergar los procesos de formación de profesionales de calidad en el país [1]De acuerdo con [2], los factores de riesgo que más influyen en la deserción estudiantil se pueden agrupar en: académicos, económicos, sociales, familiares y psicoeducativos.

De acuerdo con el Sistema para la Prevención de la Deserción en las Instituciones de Educación Superior SPADIES, con fecha de corte a marzo de 2017, el porcentaje de deserción estudiantil por cohorte fue de 45,09%. Se evidencia un porcentaje elevado si se compara con otros países como España, en el cual este porcentaje se ubica en 24,9%. Esta cifra es aún más preocupante si se compara con el promedio de deserción de la Unión Europea, el cual se ubica en el 12,8% para ese año [3].

Aunque es evidente la brecha que existe entre estos países en comparación con la situación actual en Colombia, a nivel latinoamericano esta diferencia no es tan pronunciada si se compara con países como Venezuela o Argentina, los cuales tienen un porcentaje de deserción del 52% y 43% respectivamente [3].
A pesar de que en muchos países de Latinoamérica han implementado políticas que intervengan directamente en los diferentes frentes de este fenómeno, sobre todo en los factores económicos, los índices de deserción estudiantil se mantienen en unos niveles muy altos.

Además, las instituciones de educación superior cuentan con limitados recursos y opciones para implementar estrategias de autogestión que se enfoque en cubrir gastos correspondientes a los costos crediticios, lo cual permitiría aumentar la capacidad de retención estudiantil por parte de las instituciones.

En pro de contrarrestar y prevenir los factores de riesgo que influyen directamente en la deserción estudiantil, el Ministerio de Educación Nacional implementó un sistema denominado SPADIES (Sistema para la Prevención de la Deserción de la Educación Superior). Este sistema es la herramienta para hacer seguimiento sobre las cifras de deserción de estudiantes de la educación superior. Con los datos suministrados por las instituciones de educación superior a SPADIES, se identifican y se ponderan los comportamientos, las causas, variables y riesgos determinantes para desertar. Además, con esta información se agrupan los estudiantes de acuerdo con su riesgo de deserción.

El SPADIES centraliza información proveniente de diferentes fuentes del sector que han resultado tener incidencia en la probabilidad de que un estudiante permanezca o no dentro del trayecto académico. Estas variables hacen referencia a algunas condiciones que acompañan al individuo como son su edad, género, el contexto socioeconómico que lo acompaña, la composición de su núcleo familiar, las condiciones académicas al ingresar a la educación superior y el rendimiento que obtiene durante sus estudios, entre otras [4].

1.2. FORMULACIÓN DEL PROBLEMA

La problemática planteada anteriormente generó la siguiente pregunta:

¿Qué herramienta tecnológica, basada en técnicas de Machine Learning, permitiría mejorar los niveles de deserción estudiantil a nivel de pregrado en la Universidad Autónoma del Caribe?

1.3. JUSTIFICACIÓN Y ALCANCE

La Universidad Autónoma del Caribe, como Institución de Educación Superior IES, con domicilio en la ciudad de Barranquilla, Colombia, ofrece diversos programas académicos en pro de la formación de profesionales con excelencia en busca de promover el desarrollo socioeconómico a nivel local, regional y nacional, lo que representa una exigencia constante en la revisión y renovación de todos los conceptos inherentes al ámbito competitivo y de calidad, alineado con la globalización.

Desde una perspectiva de relevancia, es importante analizar los índices de deserción estudiantil a nivel nacional. De acuerdo con el Ministerio de Educación Nacional, la tasa de deserción universitaria en el año 2010 se ubicó en un promedio del 9,89%. En el año 2015 esta tasa de deserción se mantuvo relativamente estable en comparación al año 2010, alcanzando un 9,05%.

Para el año 2018, este índice de deserción disminuyó levemente hasta ubicarse en 8,79%. Para una población de 2,2 millones de estudiantes que se matriculan anualmente en una Institución de Educación Superior en Colombia, esta cifra de deserción representa un alto costo económico y social, por lo que identificar los factores de riesgo y caracterizar a los estudiantes que se encuentran en riesgo de abandonar los estudios, ayudaría a tomar acciones preventivas y oportunas que permitan disminuir los índices de deserción.

Para lograr el cumplimiento de los objetivos, se acudió al desarrollo de un modelo analítico para la predicción de la deserción estudiantil a nivel de pregrado en la Universidad Autónoma del Caribe al cual se implementó las técnicas Machine Learning para la construcción del conjunto de datos, acompañado de herramientas como Power BI permitieron visualizar los resultados obtenidos por el modelo predictivo.

Figura 1: Factores asociados a la deserción estudiantil.

Fuente: Investigadores Luis Polo Ahumada – Darling Medina Reyes

La deserción estudiantil es una problemática social que impacta en la productividad y desarrollo de la sociedad, especialmente cuando los índices de deserción son altos y las estrategias de permanencia académica no son efectivos. De acuerdo con el programa de Permanencia Académica con Calidad y Excelencia PACE, la deserción estudiantil a nivel de pregrado en la UAC se encuentra aproximadamente en el 9% anual, lo cual se acerca al promedio nacional que es del 8,09% para el año 2020.

A nivel social, la población estudiantil que deserta a causa de un factor económico, académico, social u otro, no permite formar profesionales con excelencia académica que a futuro sean líderes regionales y nacionales en cada una de las áreas del conocimiento.

A nivel económico, la deserción estudiantil impacta en el recaudo por concepto de matrícula estudiantil de la UAC. Es por esto, que la preparación y divulgación de estrategias de permanencia académica apoyados en modelos probabilísticos, basados en comportamientos históricos de deserción, es muy importante en aras de que la UAC siga formando excelentes profesionales y mantener los niveles de acreditación y reconocimiento, que reconocen al alma máter como uno de los mejores centros educativos a nivel regional y nacional.

2. OBJETIVOS

2.1. OBJETIVO GENERAL

Desarrollar un modelo analítico que permita predecir la deserción estudiantil a nivel de pregrado en la UAC, a partir del análisis de información socioeconómica y académica de los estudiantes.

2.2. OBJETIVOS ESPECÍFICOS

- Identificar los determinantes de deserción en la Universidad Autónoma del Caribe, mediante la implementación de modelos de aprendizaje automático para la construcción del conjunto de datos a analizar.
- Implementar un modelo predictivo para predecir la probabilidad que tiene un estudiante de pregrado de la UAC de desertar de su programa académico.
- Desarrollar un dashboard en la herramienta de análisis de datos Power BI para la visualización de los resultados obtenidos en el modelo predictivo de acuerdo con estrategias establecidas en conjunto con la UAC.

3. MARCO DE REFERENCIA

3.1. ESTADO DEL ARTE

Existe un amplio espectro de investigaciones relacionadas a la planteada en este proyecto. Muchas de estas investigaciones, han sido aplicadas en diferentes Instituciones de Educación Superior dentro de sus estrategias para disminuir los índices de deserción estudiantil. A continuación, se relacionan por agrupación las categorías identificadas en los casos de estudio.

- **What Satisfies Students? Mining Student-Opinion Data with Regression and Decision Tree Analysis**: realizan el estudio en las características y experiencias que afectan la satisfacción de los estudiantes. utilizando regresión múltiple y análisis de árbol de decisión con el algoritmo del detector automático de interacción chi-cuadrado (CHAID). Los datos para este análisis provienen de una encuesta de opinión. Esta recopila datos tales como, experiencias y planes de los estudiantes; su satisfacción con el ambiente, clima, servicios e instalaciones del campus; sus percepciones del crecimiento; y las razones de su elección de universidad. Del análisis realizado por los algoritmos, hay un 93% de estudiantes que reportan un crecimiento intelectual muy grande. El autor indica que no sólo los factores académicos, sino también las variables sociales y de servicio son predictores de la satisfacción de estos estudiantes con la calidad de la educación [5].

- **Predicting students marks in Hellenic Open University:** se enfocaron en la identificación de estudiantes con bajo rendimiento académico en un sistema de aprendizaje a distancia. Analizaron variables socioeconómicas como la edad, sexo, estado civil, ocupación, número de hijos, conocimientos de informática, trabajo asociado a ordenadores y atributos académicos. Aplicaron 6 técnicas de minería, arrojando que el mejor

algoritmo fue Naıve Bayes (72,48%), seguido por el de Regresión Logística (72,32%), el de BP (72,26%) y el de SMO (72,17).

Utilizando el conjunto de datos anterior y aplicando técnicas de minería de datos y resultados obtenidos, crearon un software prototipo para clasificar las dificultades de aprendizaje de los estudiantes, permitiendo la toma de decisiones y determinando las dificultades de los estudiantes de la Universidad Abierta Helénica [6].

- **Use Data Mining to Improve Student Retention in Higher Education:** los autores argumentan cómo la minería de datos puede ayudar a detectar a los estudiantes "en riesgo", evaluar la idoneidad del curso o módulo y adaptar las intervenciones para aumentar la retención de estudiantes. Para ello, utilizaron información social de los estudiantes e información académica. Los objetivos se centraban en: Detectar patrones de Comportamiento del Estudiante, Detectar patrones de Comportamiento del Curso, Predecir de la Retención del Estudiante, Predecir de la Idoneidad del Curso, Crear una Estrategia de Intervención Personalizada. Naive Bayes obtuvo una predicción de 85.9%, las máquinas de soporte vectorial obtuvieron 78.7% y en último lugar los árboles de decisiones con 71.2% [7].

- **Study of factors analysis affecting academic achievement of undergraduate students in international program:** analizan los factores que afectan el logro académico que contribuye a la predicción del desempeño académico de los estudiantes. Útil para identificar a los estudiantes débiles que tienen probabilidades de tener un desempeño deficiente en sus estudios. El investigador aplicó el conjunto de datos para diferenciar los clasificadores (árbol de decisión, red neuronal). Se utilizó una validación cruzada con 10 pliegues para evaluar la precisión de la predicción. Los resultados muestran que el clasificador del árbol de decisión alcanza una alta precisión del 85,188%, que es superior en un 1,313% a la del clasificador de la red neuronal [8].

- **A case study of knowledge discovery on academic achievement, student desertion and student retention:** presentaron un estudio sobre el descubrimiento de conocimientos basado en el análisis de datos académicos. Los objetivos fueron obtener conocimiento sobre el éxito y el fracaso académico, la retención y la deserción estudiantil. Se utilizaron técnicas de minería de datos de clustering automático y reglas de decisión. La aplicación del algoritmo C-mean a subconjuntos de datos estadísticamente homogéneos proporcionó un grupo de conglomerados, que se han descrito cualitativamente. Utilizando clústeres seleccionados, un estudio de reglas de decisión basado en el algoritmo C4.5 generó un conjunto de reglas de decisión para los cuatro temas de investigación del estudio. Como la facultad, el programa académico, el género, la categoría de estudiante, el área de origen, etc. También contiene los datos de los estudiantes datos sobre el rendimiento académico, como la nota media académica acumulada y la nota de la prueba nacional preuniversitaria [9].

- **Predicting who will drop out of nursing courses:** A machine learning exercise: utilizaron arboles de decisiones para predecir la deserción de estudiantes de enfermería de una Universidad Británica. Empleando información socioeconómica y académica como: edad, sexo, notas del estudiante, promedio de los semestres, entre otros. Recopilando datos de 528 estudiantes en 5 años. Utilizando 3978 registros únicos, divididos en un conjunto de entrenamiento y un conjunto de pruebas. Obtuvieron una sensibilidad del 84%, especificidad del 70% y una precisión del 94%. Los autores argumentan la necesidad de tener grandes cantidades de datos y de alta calidad [10].

- **Predicting student attrition with data mining methods Journal of College Student Retention:** usando 8 años de datos institucionales junto con tres técnicas populares de minería de datos, el autor desarrolló

modelos analíticos para predecir la deserción de los estudiantes de primer año. De los tres tipos de modelos (redes neuronales artificiales, árboles de decisión y regresión logística), las redes neuronales artificiales se desempeñaron mejor, con una precisión de predicción general del 81% en la muestra de retención. El análisis de importancia variable de los modelos reveló que las variables educativas y financieras son las más importantes entre los predictores utilizados en este estudio [11].

- **Early dropout prediction using data mining: A case study with high school students:** los autores aplicaron técnicas de minería de datos para predecir el fracaso y la deserción escolares. Utilizaron datos reales de 670 estudiantes de la Universidad Autónoma de Zacatecas (UAPUAZ) para el año académico 2009/10, y empleando métodos de clasificación, tales como reglas de inducción y árboles de decisión. Los experimentos intentan mejorar su precisión para predecir qué estudiantes podrían fracasar o abandonar los estudios, primero, utilizando todos los atributos disponibles (77); luego, seleccionando los mejores atributos (15); y finalmente, reequilibrando los datos y utilizando una clasificación sensible a los costos. La precisión más alta la obtuvieron los árboles de decisión con una tasa de aciertos de 96.6% [12].

- **Extraction student dropout patterns with data mining techniques in undergraduate programs:** presentan los de un proyecto de investigación que busca identificar patrones de deserción escolar a partir de datos socioeconómicos, académicos, disciplinarios e institucionales de estudiantes de pregrado de la Universidad de Nariño de la ciudad de Pasto (Colombia), utilizando técnicas de minería de datos. Crearon un conjunto de datos con los registros de los estudiantes que fueron admitidos en los períodos comprendidos entre el primer semestre de 2004 y el segundo semestre de 2006. Se analizaron tres cohortes completas con un período de observación de seis años hasta 2011. Se descubrieron los perfiles

socioeconómicos y académicos de los estudiantes que abandonaron la escuela utilizando técnicas de clasificación basadas en árboles de decisión. Logrando una precisión superior a 80% [13].

Los estudios citados resaltan la importancia de abordar la problemática a partir de la identificación de variables académicas, socioeconómicas e institucionales que permitan identificar los factores de riesgo más relevantes que llevan a estudiantes a desertar de su programa académico. Las investigaciones se enfocan en una población específica ya sea local o regional, lo cual brinda características diferentes entre cada estudio.

Las estudios citados proponen desarrollos de modelos predictivos a partir de la minería de datos, lo cual tiene como objetivo descubrir patrones en los datos que antes de implementar el modelo eran desconocidos, sim embargo, el modelo de Machine Learning desarrollado para el presente proyecto de grado reproduce patrones de datos conocidos y realiza las predicciones con base en estos patrones.

3.2. MARCO TEÓRICO

El presente proyecto de investigación se basa en los factores conceptuales de deserción en Instituciones de Educación superior y las técnicas utilizadas para la predicción en procesos de Machine Learning. Primero, se conceptualiza el fenómeno de deserción estudiantil, analizando sus causas (individual, institucional y estatal). Además, se analiza el sistema SPADIES con componentes y cómo este ha contribuido a disminuir los índices de deserción estudiantil en Colombia. Asimismo, es necesario analizar las metodologías que más se utilizan en los procesos de Machine Learning y aprendizaje automático, esto con el fin de desarrollar los modelos predictivos a partir de las variables más relevantes.

3.2.1 Deserción en Instituciones de Educación Superior

Existen diversos componentes comunes que ayudan a delimitar un eje central definitorio para establecer una definición formal de la deserción estudiantil en instituciones de educación superior, ya que no existe un consenso sobre esta definición. El fenómeno de deserción puede ser visto desde diferentes perspectivas que van desde la individual hasta una visión más general donde se adquieren responsabilidades por parte del estado y las instituciones. En aras de ampliar la perspectiva que se tiene alrededor del fenómeno, se debe analizar las acciones a nivel estatal que sen han llevado a cabo para mitigar el flagelo de la deserción y los factores de riesgo que históricamente han incidido en este.

Puede entenderse como el abandono del sistema escolar por parte de los estudiantes, provocado por la combinación de factores que se generan tanto al interior del sistema como en contextos de tipo social, familiar, individual y del entorno. La tasa de deserción intra-anual solo tiene en cuenta a los alumnos que abandonan la escuela durante el año escolar, ésta se complementa con la tasa de deserción interanual que calcula aquellos que desertan al terminar el año escolar.

Según Universal [14], la deserción universitaria en Colombia coincide con la situación inestable del país y el marcado desinterés de los jóvenes por la educación superior formal. Las esperanzas y expectativas del alumnado no se ven satisfechas con los programas lectivos actuales y los jóvenes no confían en las virtudes económicas de Colombia en este sentido. Dentro de las principales causas de la deserción están: encarecimiento de los estudios, el efecto demográfico, la falta de innovación, la importancia de la visibilidad en red. Actualmente, en Colombia, 1 de cada 2 estudiantes se retira durante su carrera. Dicho problema relacionado con el sistema de educación superior se debe dimensionar y caracterizar en términos estadísticos si se tiene en cuenta que existen dos formas de medición: la primera de ellas da lugar a la denominada deserción por período o deserción anual que agrupa al conjunto de estudiantes

que, sin haberse graduado, acumulan dos semestres sin reanudar la matrícula en el programa académico.

El Sistema para la Prevención de la Deserción en las Instituciones de Educación Superior - SPADIES, reúne y clasifica información que faculta realizar una búsqueda a las circunstancias socioeconómicas y académicas de los estudiantes que se han incorporado a la educación superior en el país. De este modo, se puede permite saber la evolución y el estado del provecho y la caracterización académica de los estudiantes, esto es beneficioso para constituir los factores definitivos de la deserción, para evaluar el riesgo de deserción de los estudiantes y plantear y desarrollar las actividades de refuerzo, dirigidas a promover su continuidad y culminación. El SPADIES pertenece al Sistema Nacional de Información de la Educación Superior —SNIES y logra inferir como una parte propia de este, concentrado a la búsqueda especifica de un fenómeno de singular atractivo en el sector educativo, como es la deserción estudiantil [4]. El seguimiento que se realiza en el SPADIES permite conocer los siguientes aspectos [4]

- Cantidad de desertores y estimaciones de deserción
- El rasgo de cada estudiante: cualidades socioeconómicas, académicas individuales.
- Las razones o componentes claves de la deserción.
- Los datos para la valoración de rendimiento y retroalimentación de actividades llevadas a cabo para reducir la deserción estudiantil.
- Valoración del riesgo de deserción de los estudiantes.

3.2.2 Aprendizaje automático

El aprendizaje automático (ML) es el proceso mediante el cual se usan modelos matemáticos de datos para ayudar a un equipo a aprender sin instrucciones directas. Se considera un subconjunto de la inteligencia artificial (IA). El aprendizaje automático usa algoritmos para identificar patrones en los datos, y

esos patrones luego se usan para crear un modelo de datos que puede hacer predicciones. Con más experiencia y datos, los resultados del aprendizaje automático son más precisos, de forma muy similar a cómo los humanos mejoran con más práctica.

La adaptabilidad del aprendizaje automático lo convierte en una excelente opción en escenarios en los que los datos siempre cambian, la naturaleza de la solicitud o la tarea siempre se transforma o la codificación de una solución sería realmente imposible.

- ✓ **Relación del aprendizaje automático con la IA**

  El aprendizaje automático se considera un subconjunto de la IA. Un equipo "inteligente" piensa como una persona y realiza tareas por sí mismo. Una manera de entrenar un equipo para imitar el razonamiento humano es usar una red neuronal, que es una serie de algoritmos que se modelan a partir del cerebro humano.

- ✓ **Relación del aprendizaje automático con el análisis predictivo**

  Aunque el aprendizaje automático es un tipo de análisis predictivo, un gran matiz es que el aprendizaje automático es mucho más fácil de implementar con actualizaciones en tiempo real a medida que obtiene más datos. El análisis predictivo suele funcionar con un conjunto de datos estático y se debe actualizar la pantalla para ver las actualizaciones.

3.2.3 Tipos de aprendizaje automático

- Aprendizaje supervisado: aborda los conjuntos de datos con etiquetas o estructura sirve como un profesor y "entrena" al equipo, lo que aumenta su capacidad para realizar una predicción o tomar una decisión.

- Aprendizaje no supervisado: aborda los conjuntos de datos sin etiquetas ni estructuras, buscar patrones y relaciones mediante la agrupación de datos en clústeres.

3.2.4 Cómo funciona al aprendizaje automático

Paso 1: Recopilar y preparar los datos

Una vez que se identifican los orígenes de datos, se compilan los datos disponibles. El tipo de datos que tiene puede ayudar a determinar los algoritmos de aprendizaje automático que puede usar. Al revisar los datos, se identifican las anomalías, se desarrolla la estructura y se resuelven los problemas de integridad de los datos.

Paso 2: Entrenar el modelo

Los datos preparados se dividen en dos grupos: el conjunto de entrenamiento y el conjunto de pruebas. El conjunto de entrenamiento está formado por una gran parte de los datos que se usan para ajustar los modelos de aprendizaje automático con la máxima precisión.

Paso 3: Validar el modelo

Cuando esté listo para seleccionar el modelo de datos final, se usa el conjunto de pruebas para evaluar el rendimiento y la precisión.

Paso 4: Interpretar los resultados

Revise el resultado para buscar información, sacar conclusiones y predecir los resultados.

4. PROCEDIMIENTO METODOLÓGICO

4.1. METODOLOGÍA

La metodología que se utilizó en este proyecto de investigación se basa en el desarrollo de un modelo analítico predictivo usando técnicas de Machine Learning

para la predicción de la deserción de estudiantes de pregrado de la UAC. Las principales etapas del método que será aplicado en esta investigación son descritas detalladamente a continuación:

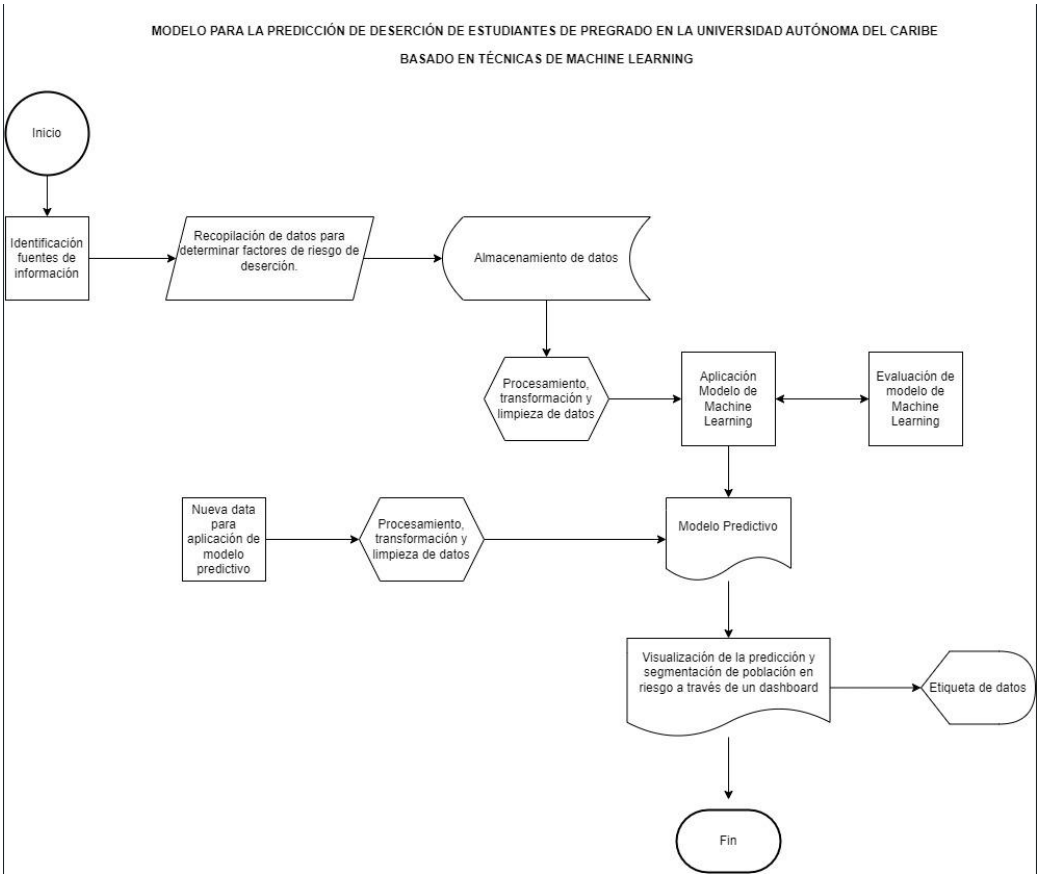

Figura 2: Modelo para la predicción de deserción de estudiantes de pregrado en la Universidad Autónoma del Caribe Basado en técnicas de Machine Learning.

Fuente: Investigadores Luis Polo Ahumada – Darling Medina Reyes

- Recopilación de datos: mediante esta etapa se obtiene la información disponible sobre los estudiantes de pregrado de la Universidad Autónoma del Caribe. Para ello, el conjunto de factores que pueden afectar el rendimiento académico de los estudiantes debe ser identificado y tomado de las diferentes fuentes de datos disponibles. Toda la información que se obtenga debe integrarse en un conjunto de datos. Las fuentes de información identificadas para la recopilación de los datos son las bases de datos institucionales de la UAC, la cual posee las variables necesarias para la implementación del modelo predictivo.

- Reprocesamiento de datos: el conjunto de datos en esta etapa está preparado para aplicar las técnicas de Machine Learning. Aplicando métodos de reprocesamiento, como limpieza de datos, transformación de variables, selección de atributos y rebalanceo de los datos. Estos últimos para resolver los problemas de alta dimensión y desequilibrio de los datos que suelen presentarse en los conjuntos de datos.

- Desarrollo modelo predictivo: en esta fase, se utilizan modelos de Machine Learning para predecir el fracaso de los estudiantes como un problema de clasificación. Se ejecutan, evalúan y comparan diferentes algoritmos para determinar cuál de ellos obtiene los mejores resultados. Se analizan los modelos obtenidos para detectar la deserción estudiantil. El desarrollo del modelo predictivo se basará en lenguaje Python o R, de acuerdo con el análisis y caracterización de los diferentes modelos de aprendizaje automático.

- Visualización de información: en esta etapa se desarrolla un dashboard en la herramienta de análisis de datos Power BI. Mediante esta herramienta se logra visualizar información relevante e indicadores clave que permitan analizar los resultados obtenidos por el modelo predictivo y así apoyar en la aplicación de estrategias de acuerdo con cada uno de los factores de riesgo detectados en la deserción estudiantil.

4.2. TIPO DE ESTUDIO

El proyecto de grado se enfoca en abordar la problemática de la deserción estudiantil en el nivel de pregrado. Para lograr este objetivo, se requiere de un estudio riguroso que permita identificar las causas subyacentes y las variables asociadas con la deserción estudiantil.

En este caso, se aplicó un estudio de investigación cuantitativa. Este tipo de estudio es adecuado para recopilar y analizar datos numéricos con el fin de responder a la pregunta de investigación planteada.

Para llevar a cabo un estudio cuantitativo sobre la deserción estudiantil en el pregrado, se pueden utilizar diversas técnicas de recolección de datos, tales como, análisis de registros académicos. Por ejemplo, se podrían analizar los registros académicos de los estudiantes para identificar patrones de deserción y correlacionarlos con otras variables.

Una vez que se han recopilado los datos, se pueden utilizar técnicas estadísticas para analizarlos y determinar las variables más importantes que están relacionadas con la deserción estudiantil. Por ejemplo, se puede utilizar un análisis de regresión logística para identificar las variables que predicen la probabilidad de que un estudiante abandone sus estudios. También se pueden utilizar técnicas de análisis de supervivencia para medir la probabilidad de que un estudiante abandone sus estudios en función del tiempo que ha estado matriculado en el programa.

Con base en los resultados obtenidos a través de estas técnicas, se puede desarrollar un modelo analítico que permita predecir la deserción estudiantil en el pregrado. Este modelo puede ser utilizado para diseñar estrategias y políticas para reducir la deserción estudiantil y mejorar la retención de los estudiantes.

4.3. CRONOGRAMA – PLAN DE TRABAJO

El siguiente plan de trabajo está conformado por el desarrollo del proyecto, y cada una de las actividades que nos sirvieron para el cumplimento de los objetivos.

Tabla 1: Cronograma de actividades.

Fuente: Grupo investigador

UNIVERSIDAD AUTONOMA DEL CARIBE
PROYECTO DE GRADO
CRONOGRAMA

COMPONENTES	DESCRIPCION	Fecha de inicio	Fecha final	Duracion (Dias)
FASE	**DURACIÓN TOTAL DEL PROYECTO**	**8/08/2022**	**15/11/2022**	95
OBJETIVO 1	**Identificar y documentar las variables socioeconómicas y académicas de los estudiantes, y las técnicas de Machine Learning, como apoyo a la construcción del conjunto de datos.**	**8/08/2022**	**10/09/2022**	33
Entregable # 1:	**Identificación y documentación de variables y técnicas de Machine Learning**	**8/08/2022**	**10/09/2022**	33
Actividad 01	**Análisis de variables socioeconómicas y académicas relevantes en la deserción estudiantil de acuerdo con estudios realizados**	**8/08/2022**	**24/08/2022**	17
Tarea 1	Identificación de factores de riesgo	8/08/2022	16/08/2022	9
Tarea 2	Caracterización de riesgos materializados	17/08/2022	24/08/2022	8
Actividad 02	**Identificación y análisis de las técnicas de Machine Learning**	**25/08/2022**	**10/09/2022**	16
Tarea 1	Documentación de técnicas de aprendizaje automático	25/08/2022	1/09/2022	8
Tarea 2	Selección de técnica de Machine Learning a utilizar de acuerdo con la caracterización de datos	2/09/2022	10/09/2022	8
OBJETIVO 2	**Implementar un modelo predictivo para predecir la probabilidad que tiene un estudiante de pregrado de la UAC de desertar de su programa académico.**	**11/09/2022**	**14/10/2022**	33
Entregable # 2:	**Desarrollo e implementación de modelo predictivo**	**11/09/2022**	**14/10/2022**	33
Actividad 01	**Desarrollo modelo predictivo**	**11/09/2022**	**27/09/2022**	17
Actividad 02	**Implementación modelo predictivo**	**28/09/2022**	**14/10/2022**	16
OBJETIVO 3	Desarrollar un dashboard en la herramienta de análisis de datos Power BI para la visualización de los resultados obtenidos en el modelo predictivo de acuerdo con estrategias establecidas en conjunto con la UAC.	**15/10/2022**	**15/11/2022**	31
Entregable # 3:	**Desarrollo de un dashboard para visualización y análisis de datos**	**15/10/2022**	**31/10/2022**	16
Actividad 01	**Definición mockups para visualización**	**15/10/2022**	**22/10/2022**	8
Tarea 1	Reunión con PACE para definición de información y objetos visuales a implementar	15/10/2022	22/10/2022	8
Actividad 02	**Configuración y desarrollo dashboard**	**23/10/2022**	**31/10/2022**	8
Tarea 1	Configuración de conexiones desde la fuente de información	23/10/2022	26/10/2022	4
Tarea 2	Desarrollo de dashboard	27/10/2022	31/10/2022	4
Entregable # 4:	**Evaluación de modelo predictivo**	**1/11/2022**	**15/11/2022**	15
Actividad 01	**Validación y correción de modelo predictivo**	**1/11/2022**	**15/11/2022**	15
Tarea 1	Planteamiento de diferentes escenarios de validación.	1/11/2022	8/11/2022	8
Tarea 2	Ajustes sobre el modelo predictivo	9/11/2022	15/11/2022	7

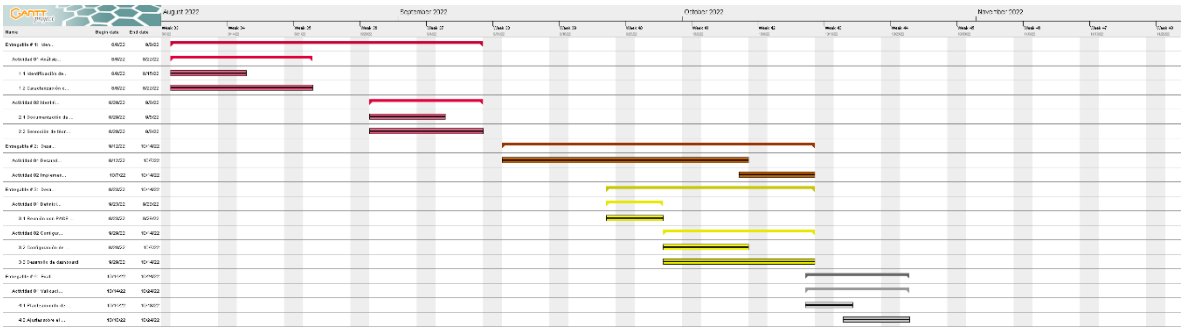

Figura 3: Gantt

Fuente: Grupo Investigador

5. PRESUPUESTO

A continuación, se muestra el presupuesto general del proyecto titulado “Modelo analítico para la predicción de la deserción estudiantil a nivel de pregrado en la Universidad Autónoma del Caribe” el cual está formado por diferentes ítems que permiten observar los costos del proyecto en diferentes áreas.,

5.1. PRESUPUESTO GENERAL

Tabla 2: Presupuesto general.

PRESUPUESTO GENERAL DEL PROYECTO					
RUBROS	**Fuentes de Financiamiento**				**Total**
	Dirección de Investigación y Transferencia	**Facultad / Programa**	**Otras fuentes Externas**	**Contrapartida UAC**	
1. Personal Científico	$ 0	$ 0	$ 0	$ 4.286.609	$ 4.286.609
2. Personal de Apoyo	$ 0	$ 0	$ 0	$ 3.569.600	$ 3.569.600
3. Consultoría Especializada y Servicios Técnicos	$ 0	$ 0	$ 0	$ 0	$ 0
4. Materiales e Insumos	$ 0	$ 0	$ 3.100.000	$ 0	$ 0
5. Salidas de Campo	$ 0	$ 0	$ 0	$ 0	$ 0
6. Equipos	$ 0	$ 0	$ 0	$ 0	$ 0
7. Bibliografía	$ 0	$ 0	$ 0	$ 0	$ 0
8. Difusión de Resultados	$ 0	$ 0	$ 0	$ 0	$ 0
9. Viajes	$ 0	$ 0	$ 0	$ 0	$ 0
TOTAL, PRESUPUESTO DEL PROYECTO	**$ 0**	**$ 0**	**$ 3.100.000**	**$ 7.856.209**	**$ 10.956.209**

5.2. PERSONAL CIENTÍFICO Y DE APOYO

El presupuesto invertido en este rubro consiste en el costo del tiempo empleado por el personal de investigación vinculados a este proyecto, que incluye a los directores y a los auxiliares de investigación.

Tabla 3: Costo personal científico.

1. PERSONAL CIENTIFICO										
Nombres y Apellidos	Función dentro del Proyecto	Tipo de Contrato	Valor Hora ($)	Dedicación Horas/semana	No. de Semanas	Fuentes de Financiamiento				
						Vicerrectoría de Investigación y Transferencia	Facultad / Programa	Otras Fuentes Externas	Contrapartida UAC	SUB-TOTAL
Jair Villanueva	Investigador Principal	Titular	$ 46.666	2	32				$ 2.986.624	$ 2.986.624
Carlos Díaz	Coinvestigador	Asociado	$ 41.935	1	31				$ 1.299.985	$ 1.299.985
					SUB-TOTAL	***$ 0***	***$ 0***	***$ 0***	***$ 4.286.609***	***$ 4.286.609***

Tabla 4. Costo personal de apoyo.

2. PERSONAL DE APOYO										
Nombres y Apellidos	Función dentro del Proyecto	Tipo de Vinculación	Valor Hora ($)	Dedicación Horas/semana	No. de Semanas	Fuentes de Financiamiento				
						Vicerrectoría de Investigación y Transferencia	Facultad / Programa	Otras Fuentes Externas	Contrapartida UAC	SUB-TOTAL
Luis José Polo Ahumada	Aux. Investigación	Practicante	$ 2.231	25	32				$ 1.784.800	$ 1.784.800
Darling Johana Medina Reyes	Aux. Investigación	Practicante	$ 2.231	25	32				$ 1.784.800	$ 1.784.800
					SUB-TOTAL	***$ 0***	***$ 0***	***$ 0***	***$ 3.569.600***	***$ 3.569.600***

5.3. CONSULTORIA ESPECIALIZADA

Tabla 5. Costo consultoría especializada.

3. CONSULTORIA ESPECIALIZADA Y SERVICIOS TECNICOS EXTERNOS				
Descripción	Justificación	Fuentes de Financiamiento		
		Vicerrectoría de Investigaciones y transferencia	INVESTIGADORES	SUB-TOTAL
1.				$ 0
2.				$ 0
3.				$ 0
	SUB-TOTAL	*$ 0*	*$ 0*	*$ 0*

5.4. MATERIALES, INSUMOS Y EQUIPOS

El presupuesto dedicado a esta sección incluye las licencias requeridas para el desarrollo del proyecto.

Tabla 6: Costo materiales e insumos. Fuente: Grupo investigador

4. MATERIALES E INSUMOS						
Descripción	Justificación	Fuentes de Financiamiento				
		Vicerrectoria de Investigación y Transferencia	Facultad / Programa	Otras Fuentes Externas	Contrapartida UAC	SUB-TOTAL
1. Licencia PRO Power BI	Será necesaria para la publicación del dashboard en el servicio en nube de Power BI				$ 452.000	$ 452.000
2.						$ 0
3.						$ 0
4.						$ 0
	SUB-TOTAL	*$ 0*	*$ 0*	*$ 0*	*$ 452.000*	*$ 452.000*

Tabla 7. Costo trabajo de campo.

5. TRABAJO DE CAMPO									
Descripción	Justificación	No. De días	No. De personas	Costo/día de estadía por persona	Transporte por persona (ida/vuelta)	Fuentes de Financiamiento			
						Vicerrectoría de Investigaciones y transferencia	INVESTIGADORES	Contrapartida UAC	SUB-TOTAL
									$ 0
									$ 0
									$ 0
SUB-TOTAL									$ 0

Tabla 8. Costo equipos usados

6. EQUIPOS						
Descripción	Justificación	Cantidad	Fuentes de Financiamiento			
			Vicerrectoría de Investigaciones y transferencia	INVESTIGADORES	Contrapartida UAC	SUB-TOTAL
1. Equipo de computo						$ 3.100.000
2.						$ 0
3.						$ 0
4.						$ 0
5.						$ 0
		SUB-TOTAL	*$ 0*	*$ 0*	*$ 0*	*$ 3.100.000*

6. PRESENTACIÓN Y ANÁLISIS DE RESULTADOS

6.1. DISEÑO DEL PROTOTIPO

El desarrollo del algoritmo se basa en diferentes fases:

Recopilación de datos

En esta fase se recolecta la información que servirá como entrada para el desarrollo del modelo, proveniente de dos fuentes: (Base de datos) proporcionada por la Universidad Autónoma del Caribe de los estudiantes que han abandonado los programas ofrecidos y (Preprocesamiento de datos) En esta fase se realiza la transformación de los datos crudos o sin procesar en un conjunto de características que se puedan utilizar como entrada para un modelo de aprendizaje automático, mediante el proceso de "featurization". El objetivo es extraer características relevantes y útiles de los datos para que el modelo pueda aprender patrones y relaciones y hacer predicciones precisas.

Desarrollo del modelo predictivo

Se ejecuta el modelo de machine learning escogido para esta predicción, que consta de un modelo lineal llamado "Regresión Logística", obtenido de la librería Sklearn. Se realizan los procesos de predicción y prueba correspondientes.

Visualización de datos

Las predicciones obtenidas en la etapa anterior se utilizan para crear un dashboard con métricas y visualizaciones aptas, mediante la herramienta de análisis PowerBI. Se anexan las KPI definidas para el análisis de este tipo de investigación con el equipo de PACE de la Universidad Autónoma del Caribe.

Como etapa final, se realiza el despliegue de este modelo, en el cual se recibe información nueva con las mismas características usadas durante el desarrollo para la ejecución de este modelo predictivo, y así obtener información real dependiendo de los diferentes factores y novedades que se reciban.

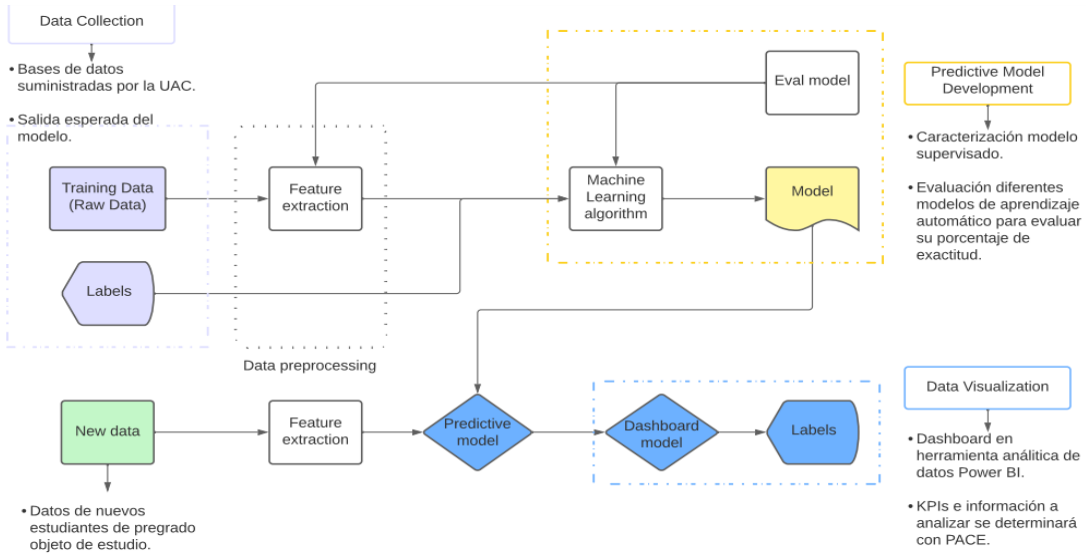

Figura 4: Adaptación de un diagrama de flujo de un modelo supervisado de Machine Learning.

Fuente: Grupo investigador

6.2. DISEÑO DISPOSITIVO FINAL

Con base en los resultados obtenidos del modelo predictivo, de plantean 3 Dashboard con información relevante que muestran el comportamiento de riesgo poblacional, probabilidades de deserción y análisis segmentado para un mayor nivel de detalle que permita focalizar la población en riesgo de deserción.

6.2.1. Dashboard: Población en Riesgo

Segmenta el grupo poblacional de estudiantes por cada facultad y presenta un análisis del comportamiento de porcentaje de repitentes por cada una de ellas. Uno de los factores claves de la deserción estudiantil es el rendimiento académico, por lo cual evaluar el comportamiento de repitentes es de suma importancia para plantear estrategias de permanencia.

La población en riesgo se segmenta a partir de los determinantes tipificados por el Ministerio de Educación Nacional, sobre los cuales la Universidad Autónoma del Caribe categoriza a cada estudiante de acuerdo con su tipología de riesgo.

En la herramienta de Inteligencia de Negocios Power BI, se desarrolló el dashboard de población en riesgo de deserción, el cual ofrece información para realizar un análisis descriptivo de la situación actual en la Universidad Autónoma del Caribe a partir de la caracterización de la población en riesgo por los determinantes establecidos por el Ministerio de Educación Nacional y un análisis por facultad de la Universidad Autónoma del Caribe el cual indica de manera porcentual la cantidad de estudiantes repitentes académicos.

En dasboard cuenta con 5 tarjetas inteligentes que, a partir de la segmentación de información que se realice, cambiará el valor del indicador mostrado. El panel de filtro está representado por un embudo en la parte superior izquierda y permite segmentar la población estudiantil por facultad, programa académico, repitentes por inasistencia o académicos, los que tiene convenios con entidades financieras, estudiantes con beca, estudiantes pertenecientes a grupos minoritarios o que cuentan con alguna discapacidad.

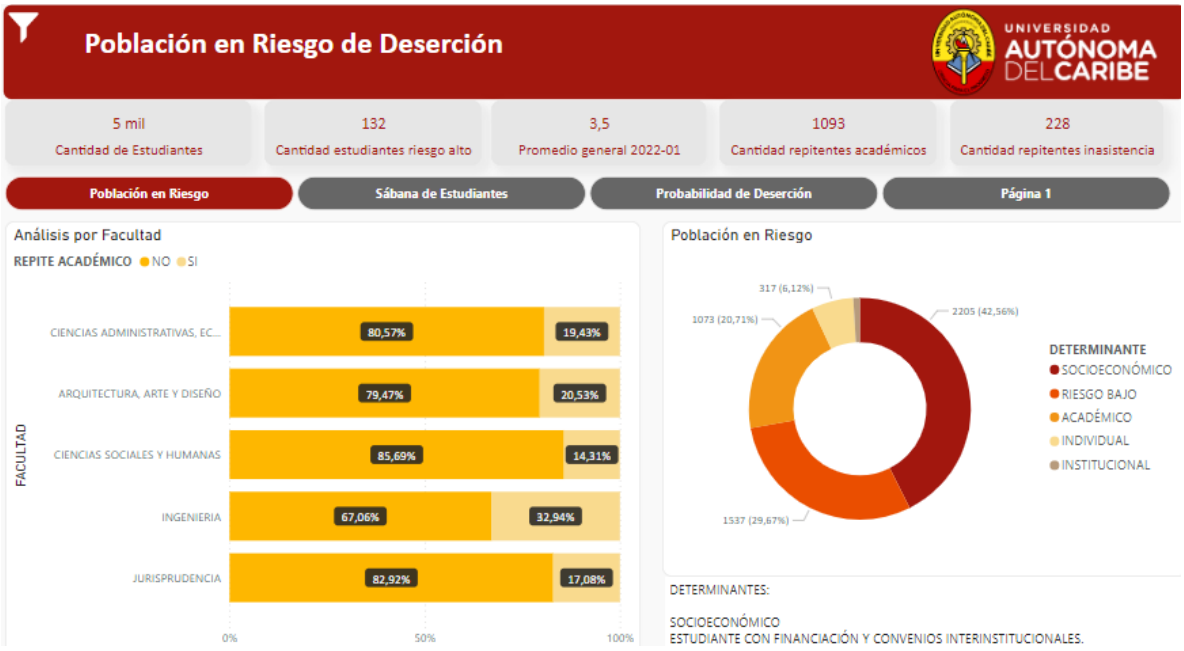

Figura 5: Dashboard (Población en riesgo)

Fuente: Grupo Investigador

6.2.2. Dashboard: Sábana de estudiantes

Detalla la probabilidad de deserción de cada uno de los estudiantes que hacen parte de la base de datos de acuerdo con los resultados obtenidos por el modelo probabilístico. El dashboard cuenta con un panel de filtros que permite visualizar de manera detallada los estudiantes por facultad, programa académico, modalidad de estudio, tipo de beca, tipo de financiación y demás segmentaciones que brindan un mayor nivel de detalle para su análisis.

La probabilidad de deserción se condiciona al nivel de porcentaje en la que esta se encuentre. Si el porcentaje de probabilidad de deserción está entre 0% y 30%, se indica de color azul, si está entre 31% y 60% se indica de color amarillo y si es mayor a 61%, este porcentaje será de color rojo. Esta categorización se realiza con el objetivo de identificar de manera visual aquellos estudiantes que mayor probabilidad de deserción tienen.

NOMBRE PROGRAMA	PERIODO	COD. ESTUDIANTE	NOTA_1ER_CORTE	NOTA_2DO_CORTE	NOTA_3ER_CORTE	NOTA_FINAL	DETERMINANTE	PREDICCION
ADMINISTRACION DE EMPRESAS	202202	10210158	2	1	3	2,0	ACADÉMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	10726191	4	5	3	4,0	ACADÉMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	10810143	3	4	2	3,0	ACADÉMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	10810404	2	3	0	2,0	ACADÉMICO	0,01 %
ADMINISTRACION DE EMPRESAS	202202	10820646	0	2	0	1,0	ACADÉMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	10911321	5	4	4	5,0	ACADÉMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	10911663	4	3	5	4,0	SOCIOECONÓMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	10920915	2	4	3	3,0	ACADÉMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	10920933	3	3	0	2,0	ACADÉMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	11010978	3	2	5	4,0	ACADÉMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	11011077	5	5	1	4,0	ACADÉMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	11011284	5	0	1	2,0	ACADÉMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	11011536	3	1	4	3,0	ACADÉMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	11016198	4	3	4	4,0	SOCIOECONÓMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	11020887	1	5	2	3,0	ACADÉMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	11111013	1	0	0	0,0	SOCIOECONÓMICO	100,00 %
ADMINISTRACION DE EMPRESAS	202202	11116656	1	5	3	3,0	ACADÉMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	11210292	5	4	5	5,0	ACADÉMICO	0,00 %
ADMINISTRACION DE EMPRESAS	202202	11211174	0	2	4	2,0	ACADÉMICO	0,00 %

Figura 6: Dashboard (Sabana de estudiantes)

Fuente: Grupo Investigador

6.2.3. Dashboard: Probabilidad de deserción

El dashboard cuenta con dos visualizaciones, la primera es una sábana de información donde se relacionan los estudiantes que tienen riesgo medio y alto de desertar de acuerdo con los resultados obtenidos a partir del modelo predictivo. Se establece como riesgo medio los estudiantes cuyo porcentaje de probabilidad de deserción se encuentra entre 21% y 60%, estos se asocian al símbolo de advertencia amarillo. Los estudiantes con riesgo alto de deserción se ubican a partir del 61% de probabilidad de desertar y se asocian al símbolo X en color rojo.

La segunda visualización representa la cantidad de estudiantes en riesgo por cada facultad de acuerdo con la regla que de clasificación que se establece para cada uno de los tipos de riesgo de deserción: bajo <= 20%, 21% < medio <= 60% y alto > 61%.

Figura 7: Dashboard (Probabilidad de deserción)

Fuente: Grupo Investigador

6.3. MATERIALES

6.3.1. Python versión 3.11:

Es el lenguaje de programación sobre el cual se desarrolló el modelo de Machine Learning. En Ciencia de Datos es uno de los lenguajes más importantes por sus capacidades de limpieza, manipulación, tratamiento y visualización de datos a partir de las librerías embebidas en este.

6.3.2. Power BI Desktop:

Es la aplicación sobre la cual se desarrolló el dashboard que permitió visualizar los resultados de la predicción del modelo de Machine Learning. Este aplicativo hace parte de la suite de Office 365 y cuenta con una amplia variedad de objetos visuales y conectores que permiten conectar a múltiples fuentes de información, para la creación de modelos relacionales, los cuales permiten la creación del dashboard que soporte la toma de decisiones.

6.3.3. Anaconda3:

Es una suite de código abierto de los lenguajes de ciencia de datos Python y R. Esta permite instalar y administrar las librerías, paquetes y entornos necesarios para el desarrollo del modelo predictivo bajo el IDE Microsoft Visual Studio Code. Su entorno permite gestionar de manera centralizada todos los recursos necesarios para el desarrollo del modelo de Machine Learning.

6.3.4. Jupyter Notebook:

Es un software de código abierto que permite crear y compartir documentos de código. Los cuadernos creados sobre Jupyter permite integrarse con el IDE Microsoft Visual Studio Code, lo cual ofrece ventajas en el desarrollo de software.

6.4. RECOLECCIÓN DE DATOS

La recolección de datos se realizó en base a la muestra seleccionada la cual se presenta a continuación:

6.4.1. MUESTRA POBLACIONAL

De acuerdo con la metodología de la investigación, la muestra poblacional se define a partir de los registros históricos de estudiantes matriculados a nivel de pregrado en la Universidad Autónoma del Caribe. Estos registros hacen parte del periodo académico 2022-01, lo cual brinda una visión de la estructura y tipo de datos con los cuales se desarrollará el modelo predictivo.

Los datos son extraídos directamente de la base de datos proporcionados por el programa de Permanencia Académica con Calidad y Excelencia PACE, los cuales componen el mundo poblacional de estudiantes de pregrado de todas las facultades de la UAC.

Dado el enfoque de la investigación, orientada a plantear nuevas herramientas que permitan fortalecer la permanencia académica, todos los estudiantes de pregrado de la UAC identificados y consolidados por el PACE, son el grupo poblacional idóneo para el desarrollo del modelo predictivo.

6.5. ANÁLISIS DE RESULTADOS

La toma de decisiones a partir de resultados obtenidos por modelos probabilísticos permite dar respuestas a necesidades o problemáticas sociales o de negocio.

En esta investigación, se aplicó un modelo de regresión logística, la cual es una técnica de machine learning que permite predecir el resultado de una variable categórica en función de unas variables independientes o predictores.

6.5.1. ANÁLISIS DE LAS PRUEBAS REALIZADAS POR EL DISPOSITIVO FINAL

Observamos inicialmente cómo lucia la base de datos de forma descriptiva,

	PERIODO	PERIODO INGRESO	COD. ESTUDIANTE	PROMEDIO 2022-01	NOTA_1ER_CORTE	NOTA_2DO_CORTE	NOTA_3ER_CORTE	NOTA_FINAL	PUNTAJE SABER 11
count	5181.0	5178.000000	5.181000e+03	4299.000000	5181.000000	5181.000000	5181.000000	5181.000000	4722.000000
mean	202202.0	202080.735419	4.413638e+08	3.840986	2.495657	2.516889	2.478865	2.744876	229.414019
std	0.0	142.459363	5.335316e+08	0.726296	1.717340	1.720183	1.700723	1.082980	106.040997
min	202202.0	201401.000000	1.021016e+07	0.000000	0.000000	0.000000	0.000000	0.000000	-1.000000
25%	202202.0	202001.000000	5.211023e+07	3.600000	1.000000	1.000000	1.000000	2.000000	215.000000
50%	202202.0	202102.000000	1.618200e+08	4.000000	2.000000	3.000000	2.000000	2.700000	257.000000
75%	202202.0	202201.000000	9.516108e+08	4.300000	4.000000	4.000000	4.000000	3.500000	292.000000
max	202202.0	202202.000000	1.672220e+09	5.000000	5.000000	5.000000	5.000000	5.500000	2570.000000

Figura 8: Variables cuantitativas de la BD

Autor: Investigadores Luis Polo Ahumada – Darling Medina Reyes

Se observa que las calificaciones de los estudiantes son las únicas variables numéricas en la BD que nos interesan en nuestro modelo. También es necesario resaltar el hecho de que, todos los estudiantes son de pregrado y matriculados. Además de que existen 98 estudiantes con notas mayores que 5. Lo cual verificamos en la Figura 10 con el diagrama de caja e histograma de las notas finales. Ahora; en cuanto a las cualitativas de interés,

	TIPO INGRESO	MODALIDAD	DESCUENTO	BECA	REPITE_ACADEMICO	REPITE_INSASISTENCIA	DETERMINANTE	DESERTO
count	5178	5178	5178	5178	5178	5178	5178	5178
unique	4	2	2	2	2	2	5	2
top	POR PRIMERA VEZ	PRESENCIAL	SI	NO	NO	NO	SOCIOECONÓMICO	NO
freq	3383	5036	3542	5097	4085	4950	2202	5101

Figura 9: Variables cualitativas de la BD

Autor: Investigadores Luis Polo Ahumada – Darling Medina Reyes

Se identifica que se tratan de categorías, algunas con solamente sí o no de respuesta. Tenemos lista nuestra BD para el modelo, pero nos interesa esencialmente responder, ¿Qué factores determinan la deserción de un estudiante universitario? Sería muy intuitivo pensar que factores como la calificación final del semestre, tener beca, ser de primer semestre o incluso el determinante (académico, económico, individual, etc.) serán las variables clave de la deserción. En este articulo veremos la respuesta a esta pregunta objetivo.

- **TRANSFORMANDO LOS DATOS**

Las variables categóricas de interés en la BD las transformamos a ceros y unos con el fin de aplicar el modelo en Python de forma óptima y que la BD se adapte a las funciones del paquete sklearn con el que realizaremos el modelo. De este modo; agrupamos por promedio las variables de interés dependiendo a la deserción real estudiantil, dado que dicho dato lo conoce la universidad y nuestro objetivo es predecir por medio del modelo logístico cuándo esto se da.

DESERTO	MODALIDAD	DESCUENTO	BECA	REPITE_ACADEMICO	REPITE_INSASISTENCIA	NOTA_FINAL
0	0.972162	0.687316	0.015683	0.199177	0.042345	2.766987
1	1.000000	0.467532	0.012987	1.000000	0.155844	1.307792

Figura 10: Deserción de estudiantes

Autor: Investigadores Luis Polo Ahumada – Darling Medina Reyes

Vemos que los que desertan tienen un promedio mucho menor a los que no, hecho que es de esperarse. También observamos que la modalidad en casi todo caso es presencial, junto con la beca. La mayoría de los estudiantes no son becados, dichas variables junto con las demás son categóricas numéricas donde la deserción lo medimos en 1 si en efecto deserta o 0 si no. El código que genera dicha transformación y la agrupación es el siguiente,

```
data['MODALIDAD'] = np.where(data['MODALIDAD'] == 'PRESENCIAL', 1, 0)
data['TIPO INGRESO'] = np.where(data['TIPO INGRESO'] == 'POR PRIMERA VEZ', 0,data['TIPO INGRESO'])
data['TIPO INGRESO'] = np.where(data['TIPO INGRESO'] == 'REINGRESO', 1,data['TIPO INGRESO'])
data['TIPO INGRESO'] = np.where(data['TIPO INGRESO'] == 'TRANSFERENCIA EXTERNA', 2,data['TIPO INGRESO'])
data['TIPO INGRESO'] = np.where(data['TIPO INGRESO'] == 'TRANSFERENCIA INTERNA', 3,data['TIPO INGRESO'])
data['DESCUENTO'] = np.where(data['DESCUENTO'] == 'SI', 1, 0)
data['BECA'] = np.where(data['BECA'] == 'SI', 1, 0)
data['REPITE_ACADEMICO'] = np.where(data['REPITE_ACADEMICO'] == 'SI', 1, 0)
data['REPITE_INSASISTENCIA'] = np.where(data['REPITE_INSASISTENCIA'] == 'SI', 1, 0)
data['DETERMINANTE'] = np.where(data['DETERMINANTE'] == 'SOCIOECONÓMICO', 0,data['DETERMINANTE'])
data['DETERMINANTE'] = np.where(data['DETERMINANTE'] == 'RIESGO BAJO', 1,data['DETERMINANTE'])
data['DETERMINANTE'] = np.where(data['DETERMINANTE'] == 'ACADÉMICO', 2,data['DETERMINANTE'])
data['DETERMINANTE'] = np.where(data['DETERMINANTE'] == 'INDIVIDUAL', 3,data['DETERMINANTE'])
data['DETERMINANTE'] = np.where(data['DETERMINANTE'] == 'INSTITUCIONAL', 4,data['DETERMINANTE'])
data['DESERTO '] = np.where(data['DESERTO '] == 'SI', 1, 0)
data.groupby('DESERTO ').mean()
```

Figura 11: Código para la transformación de variables

Autor: Investigadores Luis Polo Ahumada – Darling Medina Reyes

Simplemente utilizamos la función where de numpy para establecer de forma binaria (1 o 0) si tiene o no beca, por ejemplo. Para la variable tipo de ingreso o determinante, se establecen también las categorías correspondientes de manera numérica.

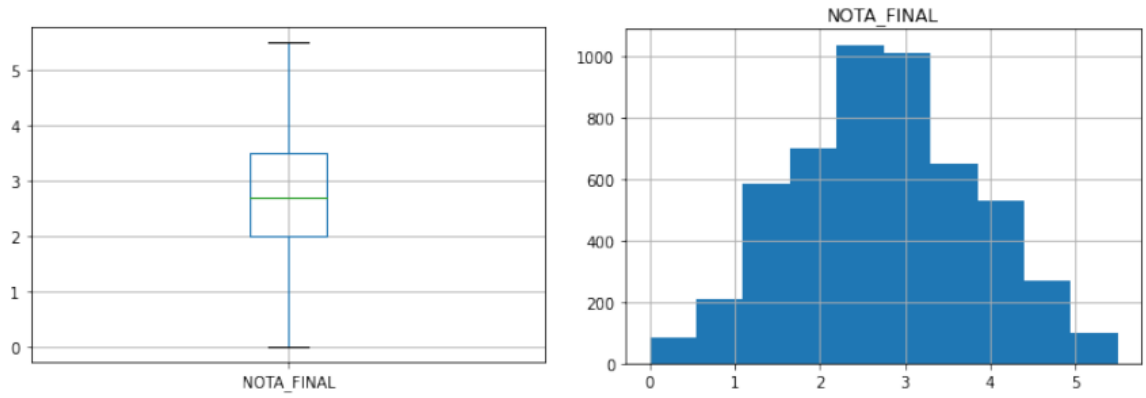

Figura 12: Frecuencia y comportamiento de las calificaciones finales.

Autor: Investigadores Luis Polo Ahumada – Darling Medina Reyes

- **VISUALIZAR Y RESUMIR LOS DATOS**

La mayoría de los estudiantes vimos que se agrupan alrededor de la nota de aprobación del semestre (3.0) donde la mediana es menor que la nota de aprobación, habiendo estudiantes con notas mayores de 5.0, lo cual desconocemos la razón. Debido a que la nota final es un promedio de las materias o las notas del corte. Lo cual, inferimos entonces que el o los docentes no cumplen con el debido intervalo de calificaciones dado por la universidad.

	MODALIDAD	DESCUENTO	BECA	REPITE_ACADEMICO	REPITE_INSASISTENCIA	NOTA_FINAL	DESERTO
MODALIDAD	1.000000	0.018149	-0.016950	-0.017462	0.036038	0.016964	0.020631
DESCUENTO	0.018149	1.000000	-0.185489	-0.144212	-0.062693	-0.001769	-0.057221
BECA	-0.016950	-0.185489	1.000000	-0.046139	-0.027055	0.014421	-0.002630
REPITE_ACADEMICO	-0.017462	-0.144212	-0.046139	1.000000	0.371080	0.001486	0.237522
REPITE_INSASISTENCIA	0.036038	-0.062693	-0.027055	0.371080	1.000000	-0.006283	0.066957
NOTA_FINAL	0.016964	-0.001769	0.014421	0.001486	-0.006283	1.000000	-0.163137
DESERTO	0.020631	-0.057221	-0.002630	0.237522	0.066957	-0.163137	1.000000

Figura 13: Correlaciones de la BD

Autor: Investigadores Luis Polo Ahumada – Darling Medina Reyes

Claramente, cada variable no está correlacionada con las demás. Solamente las que determinan si el estudiante repite por inasistencias o por temas académicos, se correlacionan con la nota final. Dado que claramente una depende de la otra. Este hecho es importante para darnos cuenta que en efecto se trata de una situación real que fácilmente se podrían presentar en ámbitos laborales en un futuro. Ahora, no hace falta verificar que la tasa de deserción por lo general es muy baja, considerando incluso si fuera del 20 %, una BD de datos desbalanceada para el modelo logístico.

En ese caso, la predicción del modelo sería que todos no desertan y claramente, no es el objetivo. Es por esto, que necesitamos del método SMOTE que consiste en lo siguiente, El método SMOTE (Synthetic Minority Oversample TEchnique) crea nuevas observaciones sintéticas en lugar de realizar un sobre muestreo por sustituciones, por medio de una interpolación lineal para la clase minoritaria. Uniendo todas las clases minoritarias que están cerca de los vecinos, de manera aleatoria. Cabe aclarar que la información precia no se pierde, el método es sencillo y fácil de interpretar, además de mejorar el overfitting.

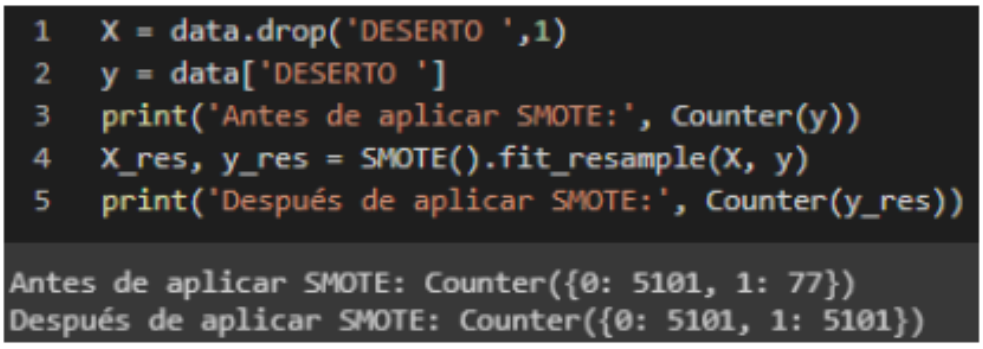

Figura 14: SMOTE en Phyton

Autor: Investigadores Luis Polo Ahumada – Darling Medina Reyes

```python
model = sk.linear_model.LogisticRegression()
model.fit(X res,y res)
X_train, X_validation, Y_train, Y_validation = sk.model_selection.train_test_split(X, y, test_size=0.2)
ns_probs = [0 for _ in range(len(Y_validation))]
lr_predicts = model.predict(data.drop(['DESERTO '],1))
lr_probs = model.predict_proba(X_validation)
lr_probs = lr_probs[:, 1]
ns_auc = roc_auc_score(Y_validation, ns_probs)
lr_auc = roc_auc_score(Y_validation, lr_probs)

print('Sin entrenar: ROC AUC=%.3f' % (ns_auc))
print('Regresión Logística: ROC AUC=%.3f' % (lr_auc),'\n')
ns_fpr, ns_tpr, _ = roc_curve(Y_validation, ns_probs)
lr_fpr, lr_tpr, _ = roc_curve(Y_validation, lr_probs)

data.loc[:,'PREDICCION'] = lr_predicts
plt.figure(figsize=(8,4)); plt.title('Curva ROC')
plt.plot(ns_fpr, ns_tpr, linestyle='--', label='Sin entrenar')
plt.plot(lr_fpr, lr_tpr, marker='.', label='Regresión Logística')
plt.xlabel('Tasa de Falsos Positivos'); plt.ylabel('Tasa de Verdaderos Positivos')
plt.legend(); plt.show()
```

Figura 15: Modelo logístico múltiple

Autor: Investigadores Luis Polo Ahumada – Darling Medina Reyes

Aplicando el modelo entonces a esta nueva BD con nuestra variable objetivo DESERTO, aplicamos el modelo con una proporción de la BD del 20 % para pruebas o validación y el 80 % restante para entrenamiento, esto con la base original. Mientras que el modelo con la base aplicando SMOTE, obteniendo así las probabilidades y la predicción de qué estudiantes desertaron. Donde finalmente, graficamos la curva ROC del modelo.

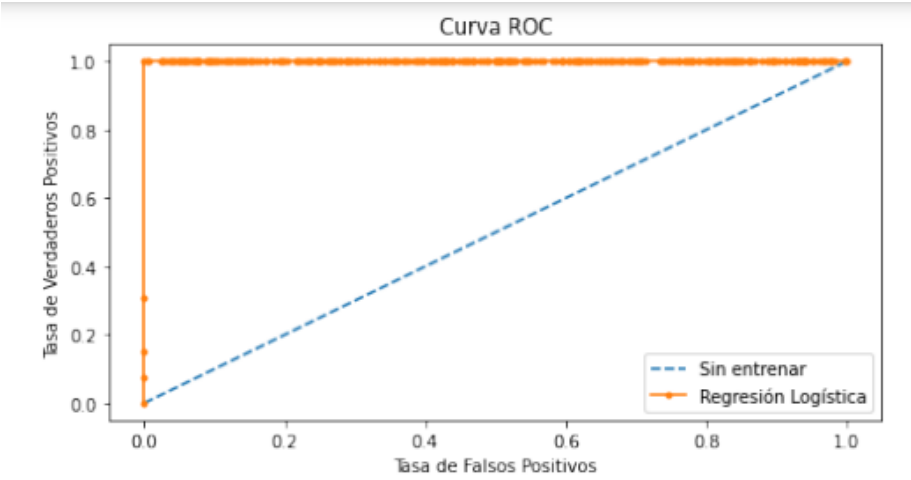

Figura 16: Curva ROC del modelo.

Autor: Investigadores Luis Polo Ahumada – Darling Medina Reyes

Lo cual está claramente relacionado con la matriz de confusión,

$$\begin{pmatrix} Predicción & 0 & 1 \\ Deserción\ real & & \\ 0 & 5029 & 72 \\ 1 & 0 & 77 \end{pmatrix}$$

Es decir, no hubo ningún falso positivo donde alguien que desertó, el algoritmo dijera que sí. Mientras que hubo 72 casos en los que el modelo predijo que sí desertaban cuando en realidad no lo hizo. Vemos que dicha cantidad es muy baja comparado con el tamaño de la muestra, por esta razón curva ROC arroja tan buenos resultados.

```
Porcentaje de desersión real:
No desertaron:  98.51 % ( 5101 )
Desertaron:  1.49 % ( 77 )
----------------------------------------
Porcentaje de desersión predicho:
No desertaron:  97.12 % ( 5029 )
Desertaron:  2.88 % ( 149 )
----------------------------------------
Porcentaje de predicciones extra:  6.49 %
```

Figura 17: Resumen del modelo

Autor: Investigadores Luis Polo Ahumada – Darling Medina Reyes

En resumen, el modelo tuvo un error de predicción de tan sólo 6,5 %, aproximadamente. Donde de los 77 estudiantes que en realidad desertaron, el modelo predijo 149. Lo cual, es buen factor para predecir qué estudiantes tuvieron tendencia a desertar, pero al final no lo hicieron. Dichos factores los resumimos en la siguiente tabla (filtrado los que no desertaron y a la vez el algoritmo predijo que sí),

	TIPO INGRESO	MODALIDAD	DESCUENTO	BECA	REPITE_ACADEMICO	REPITE_INSASISTENCIA	DETERMINANTE	NOTA_FINAL	DESERTO	PREDICCION
335	0	1	0	0	1	0	0	2.1	0	1
348	0	1	0	0	0	0	0	0.0	0	1
355	3	1	1	0	0	0	0	0.0	0	1
370	0	1	0	0	1	1	0	2.0	0	1
519	3	1	0	0	1	0	0	2.2	0	1
...	...	...	...	...	...	...	...	...	...	...
4658	0	1	0	0	1	0	0	2.2	0	1
4666	0	1	0	0	0	0	0	0.0	0	1
4693	1	1	1	0	1	1	0	2.0	0	1
4752	0	1	1	0	1	0	0	2.0	0	1
4847	0	1	1	0	1	0	0	2.2	0	1

72 rows × 10 columns

Figura 18: Variables que afectan la deserción

Autor: Investigadores Luis Polo Ahumada – Darling Medina Reyes

Podemos notar por estadística que,

1. La mayoría es de reingreso y primera vez en la universidad. (87 %)
2. Todos son presencial y no tienen beca.
3. La mitad tenían descuento en su matrícula y la otra mitad no.
4. El determinante es socioeconómico para el 98 % de los estudiantes.
5. La nota final de todos los estudiantes claramente es menor que la de aprobación con el 20 % de los estudiantes con una nota de 0, suponemos que por inasistencia.

6.6. MANUAL DE USUARIO

EL manual de usuario de este proyecto “Modelo analítico para la predicción de la deserción estudiantil a nivel de pregrado en la Universidad Autónoma del Caribe” contiene una introducción, descarga e instalación de cada uno de los programas necesarios para el desarrollo de esta, preparación del área del trabajo y preparación de los datos, cada uno de estos ítems cuenta con una serie de pasos e imágenes que serán de ayuda para su utilización.

CONCLUSIONES Y RECOMENDACIONES

Como conclusión, se logró identificar las variables relevantes que afectan directamente el fenómeno de deserción estudiantil, sus principales causas y los riesgos que se materializan con la consecución de la deserción. Esto se presenta como un reto para la sociedad, porque influye directamente en diversos componentes de esta y se traduce en una problemática social que se debe abordar.

Además, se presenta como una gran oportunidad el poder implementar modelos probabilísticos que apalanquen las estrategias implementadas en las instituciones de educación superior con el fin de garantizar la estabilidad estudiantil de acuerdo con una personalización obtenida a través de una serie de parámetros y definiciones establecidas.

Adicionalmente, se resalta la importancia de contar con la disponibilidad de las fuentes de información actualizadas y estructuradas, de tal manera que la caracterización y análisis de esta se pueda realizar en los términos requeridos para el desarrollo de los diferentes componentes del proyecto de investigación.

Cabe resaltar la importancia de implementar una arquitectura tecnológica asociada a recursos en nube como Oracle Cloud y Power BI Service que permitan automatizar el desarrollo y ejecución de desarrollos como el realizado en el presente proyecto de investigación. Estos son recursos que la Universidad Autónoma del Caribe debe licenciar si su expectativa es lograr madurar este tipo de iniciativas.

BIBLIOGRAFÍA

[1] F. Sánchez Torres y J. Márquez Zuñiga, Artists, *La deserción en la educación superior en Colombia durante la primera década del siglo XXI : ¿por qué ha aumentado tanto?.* [Art]. Universidad de los Andes, Facultad de Economía, 2012.

[2] R. Martelo, K. Herrera y N. Villabona, «Estrategias para disminuir la deserción universitaria mediante series de tiempo y multipol,» *Espacios,* p. 25, 2017.

[3] K. Cubillos Pineda, «Universidad del Rosario,» 16 Noviembre 2019. [En línea]. Available: https://www.urosario.edu.co/Periodico-NovaEtVetera/Sociedad/Desercion-en-la-Educacion-Superior-en-Latinoameric/.

[4] Ministerio de Educación Nacional, «SPADIES,» 2020. [En línea]. Available: https://www.mineducacion.gov.co/sistemasinfo/spadies/Informacion-Institucional/254648:Que-es-el-SPADIES.

[5] E. Thomas y N. Galambos, «What Satisfies Students? Mining Student-Opinion Data with Regression and Decision Tree Analysis,» *Research in Higher Education,* pp. 251-269, 2004.

[6] S. B. Kotsiantis y P. E. Pintelas, «Predicting students marks in Hellenic Open University,» *International Conference on Advanced Learning Technologies,* pp. 664-668, 2005.

[7] Y. Zhang y et al., «Use Data Mining to Improve Student Retention in Higher Education,» pp. 190-197, 2010.

[8] P. Cheewaprakobkit, «Study of factors analysis affecting academic achievement of undergraduate students in international program.,» *Lecture Notes in Engineering and Computer Science,* pp. 332-336, 2013.

[9] Salazar et al., «A case study of knowledge discovery on academic achievement, student desertion and student retention,» de *2nd International Conference on Information Technology*, 2004.

[10] L. G. Moseley y D. M. Mead, «Predicting who will drop out of nursing courses: A machine learning exercise,» *Nurse Education Today,* pp. 469-475, 2008.

[11] D. Delen, «Predicting student attrition with data mining methods,» *Journal of College Student Retention: Research, Theory and Practice,* pp. 17-35, 2011.

[12] C. e. a. Márquez-Vera, «Early dropout prediction using data mining: A case study with high school students,» *Expert Systems,* pp. 107-124, 2016.

[13] R. T. e. a. Pereira, «Extraction student dropout patterns with data mining techniques in undergraduate programs,» *5th International Conference on Knowledge Discovery and Information Retrieval and KMIS,* pp. 136-142, 2013.

[14] E. Universal, «El Botiquín.mx,» 6 Noviembre 2017. [En línea]. Available: https://www.elbotiquin.mx/medicina-general/el-color-de-tu-pipi-te-dira-si-debes-ir-al-medico. [Último acceso: 25 Mayo 2019].

[15] S. K. D. L. S. M. Strasinger, analisis de orina y de los liquidos corporales, panamericana, 2008.

[16] F. E.-C. /. J. E. Schmalbach, *facultad de medicina,* Bogotá.

[17] contaval, «contaval,» 18 02 2016. [En línea]. Available: http://www.contaval.es/que-es-la-vision-artificial-y-para-que-sirve/. [Último acceso: 10 09 2018].

[18] M. M. Julian Pérez porto, «Definición.De,» Julian Pérez porto, María Merino, 2014. [En línea]. Available: https://definicion.de/rgb/. [Último acceso: 02 08 2018].

[19] c. moler, «MATLAB,» Fabricantes de Matlab, 1984. [En línea]. Available: la.mathworks.com. [Último acceso: 4 octubre 2018].

[20] E. l. d. V. Carlos Eduardo D´Negri, «investigaciones medicas logica difusa,» buenos aires.

[21] E. G. J. H. Castillo, Systems and probabilistic Network Models, Springer Verlag, New York: Castellana, 1998.

[22] M. A. G. German Campuzano Maya, «Uroanalisis: Mas que un examen de rutina,» editora medica colombiana, Antioquia, 2006.

[23] M. I. R. e. e. D. M. y. V. David Saceda Corralo, «Webconsultas,» 2017 julio 2017. [En línea]. Available: https://www.webconsultas.com/pruebas-medicas/sedimento-urinario. [Último acceso: 30 mayo 2019].

[24] F. Rodriguez, «Blog de Laboratorio Clínico y Biomédico,» 2 agosto 2017. [En línea]. Available: https://www.franrzmn.com/analisis-fisico-quimico-de-la-orina/. [Último acceso: 30 mayo 2019].

[25] M. Yamini Durani, «Wake Forest,» febrero 2012. [En línea]. Available: https://www.brennerchildrens.org/KidsHealth/Parents/Cerebral-Palsy-Center/En-espanol/Analisis-de-la-orina-tiras-reactivas.htm. [Último acceso: 30 mayo 2019].

[26] Vistronica, «Vistronica,» [En línea]. Available: https://www.vistronica.com/domotica/camara-usb-hd-720p-detail.html. [Último acceso: 30 mayo 2019].

[27] Arkray, «Arkray,» [En línea]. Available: http://www.arkraylatam.com/spanish/products/laboratory/test_strips/aution_sticks_10ea.html. [Último acceso: 30 mayo 2019].

[28] Maker, «Maker,» [En línea]. Available: https://somosmaker.com/producto/pla_blanco/.

] [Último acceso: 30 mayo 2019].

[29 «Tiras de Iluminación,» 22 febrero 2013. [En línea]. Available:
] https://tirasledIluminacion.com.mx/blog/noticias/que-son-las-tiras-led-y-como-funcionan/. [Último acceso: 30 mayo 2019].

[30 M. G. P. A. C. C. G. E. N. F. P., «Diagnóstico Clínico de la Hematuria Vesical Enzoótica Bovina
] por Urianálisis». Perú 11 marzo 2017.

[31 F. P. V. M. G. C. R. J. F. Sofía Di Chiazza, «Análisis de orina: estandarización y control de
] calidad». 21 febrero 2014.

[32 G. F. F. C.-B. A. P. A. R. M. GOMILA MUNOZ ISABEL, «DISPOSITIVO PORTATIL DE ANALISIS DEL
] PH DE LA ORINA». 8 marzo 2010.

[33 J. P. A. ROBERT FRANCIS EISELE, « DISPOSITIVO PARA SUJETAR UNA TIRA REACTIVA PARA
] ANÁLISIS DE LÍQUIDOS». 23 mayo 2018.

[34 H. I. G. Said David Pertuz Arroyo, «Sistema de adquisición automática de imágenes para
] microscopio óptico». 30 octubre 2007.

[35 K. D. Desai, «sistema automatizado para el análisis de orina: un método simple, rentable y
] confiable para distinguir entre fuentes glomerulares y no glomerulares de hematuria». 25 Diciembre 2001.

[36 C. d. l. U. G. Mason, «Desarrollan un test para diagnosticar tuberculosis a partir de la orina».
] 19 Diciembre 2017.

[37 Nefrotest, «Desarrollo de una aplicación capaz de realizar la lectura de las tiras reactivas».
] Colombia 2018.

[38 M. Augstein y S. y. S. R. Riebel, «Dispositivo de diagnóstico rápido con bloqueo mediante tiras
] reactivas». 13 Octubre 2010.

[39 G. YOUNG, M. O'CONNELL, I. MCARTHUR, A. MCNEILAGE y N. y. A.-I. M. PHIPPEN, «Tintas
] reactivas enzimáticas para su uso en tiras reactivas que tiene un código de calibración predeterminado». 29 Junio 2015.

[40 D. M. Koon-Wah Leong, «Sistema de recipiente para tiras reactivas.». 3 Junio 2002.
]

[41 J. y. W. R. E. LOVELL, «Método para dispensar tiras reactivas para diagnóstico». 27 Junio 2014.
]

[42 J. y. W. R. E. LOVELL, «Método para dispensar tiras reactivas para diagnóstico». 27 Junio 2014.
]

[43] D. M. Koon-Wah Leong, «Sistema de recipiente para tiras reactivas.». 3 Junio 2002.

[44] H. LEE y M. A. y. H. H. Y. DINEVA, «Interacciones de unión mejorada en ensayos con tiras reactivas.». 20 Enero 2012.

[45] K. CLAUSEN, «configuración de un dispositivo de tira reactiva seca y procedimiento para determinar un analítico en una muestra utilizando dicho dispositivo de tira reactiva seca». 2012 Junio 2012.

[46] V. y. R. S. S. QUIRELL JOSE, «Procedimiento y sistema de medición de una tira reactiva». 19 Abril 2013.

[47] A. P. Phelan, «Disposición óptica para dispositivo de lectura de análisis». 13 Julio 2009.

[48] J. MONDRO, «Tira reactiva de diagnóstico que tiene características de transporte de fluido». 14 Junio 2017.

[49] J. CREMINS, «Método y composición para teñir y procesar una muestra de orina». 12 Marzo 2018.

[50] D. Hessels, G. Verhaegh y J. A. y. W. A. J. Schalken, «Razones de arnm en sedimentos urinarios y/o orina como pronostico y/o marcador para el tratamiento y el diagnóstico de cáncer de prostata». 4 Junio 2010.

[51] A. BERGMANN, «Método para diagnosticar o monitorizar la función renal o diagnosticar la disfunción renal». 4 Julio 2018.

[52] M. Peña Cabrera, I. López Juárez, H. Gómez N., R. Osorio C. y O. Sergiyenko.Mexico 2009.

[53] M. Peña Cabrera, I. López Juárez, H. Gómez N., R. Osorio C. y O. Sergiyenko, «Automatización del proceso de ensamble utilizando visión artificial». Mexico 2009.

[54] W. E. H. R. H. G. E. R. David Hough, «Aparato espectrofotométrico con detección de tiras reactivas.». 1999.

[55] I. Willis E. Howard, «Sistema de reconocimiento óptico de códigos sobre una tira de pruebas de diagnóstico.». 16 Noviembre 2006.

[56] Y. R. K. R. W. L. P. E. L. J. P. S. Vernon L. Chupp, «Procedimiento y aparato para la realizacion de analisis automatizados.». 1 junio 2006.

[57] N. A. S. T. F. R. K. R. P. D. W. Raghbir Singh Bhullar, «Tira reactiva con cámara de recepción de muestra ensanchada». 9 Diciembre 2018.

[58] J. A. S. G. V. A. J. W. Daphne Hessels, «Razones de arnm en sedimentos urinarios y/o orina

] como pronostico y/o marcador para el tratamiento y el diagnóstico de cáncer de próstata.». 4 Junio 2010.

[59] «INNOVAR TECNOLOGÍA BIOMÉDICA S.A.S.,» 2012. [En línea]. Available: https://www.innovar.com.co/shop/category/laboratorio-centrifugas-34. [Último acceso: 25 05 2019].

[60] M. E. Cancino, «Visita y Salud,» [En línea]. Available: http://tusanascondios.blogspot.com/2015/10/que-significado-tiene-el-color-de-mi.html. [Último acceso: 25 Mayo 2019].

[61] «Probak BC,» [En línea]. Available: https://www.probakbc.com/producto/frasco-esteril-para-colectar-muestra-de-orina/. [Último acceso: 25 Mayo 2019].

[62] K. García Mendoza y S. Soto Cantero, Artists, *Factores determinantes de la deserción tardía y la graduación en la Universidad de la Costa CUC.* [Art]. Universidad de la Costa.